بنام خدا

الگوشناسی آماری (CE-725)

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف

تمرینات سری ششم – بهار ۱۳۸۹

به نکات زیر توجه فرمائید:

۱. زمان تحویل تمرینات در سایت درس مشخص شده است. دقت نمائید که زمانبندی‌های تعیین شده قابل تغییر نیستند.

۲. تمرینات را با عنوان SPR-HWx-8xxxxxxx (مثلا SPR-HW6-88300785) و در یک فایل فشرده با همین نام به آدرس Muhammadi@ce.sharif.edu ایمیل زده و *در اولین جلسه بعد از زمان تحویل، بصورت پرینت شده تحویل استاد درس دهید.*

۳. گزارش شما باید مختصر و مفید باشد. برای تمرینات پیاده‌سازی که با لوگوی MATLAB مشخص شده‌اند باید کد مطلب نوشته شده ضمیمه گزارش شده و تمامی خروجی‌های برنامه‌ها در گزارش شما ذکر شوند.

سوال ۱) دو داده زیر را از دو کلاس ۱+ و ۱− در نظر بگیرید:

$x_1 = (0, 0), y_1 = +1$

$x_2 = (4, 4), y_2 = -1$

الف) می‌خواهیم یک کلاسه‌بند SVM خطی برای این داده‌ها طراحی کنیم. ابتدا بردارهای پشتیبان را مشخص کرده و سپس با توجه به آنها کلاسه‌بند زیر را کامل کنید:

$$h(x) = \begin{cases} +1 & if \ \ .....x1 + ...x2 + ... \geq 0 \\ -1 & otherwise \end{cases}$$

ب) فرض کنید داده زیر نیز به داده‌های قبلی اضافه شود:

$x_3 = (-1,-1), y_3 = +1$

همانند حالت قبل بردارهای پشتیبان را مشخص کرده و مرز تصمیم‌گیری را بیابید. آیا مرز تصمیم‌گیری برای این حالت متفاوت از حالت (الف) خواهد بود؟ وزن بردارهای پشتیبان در این حالت نسبت به حالت قبلی کوچک‌ترند یا بزرگ‌تر؟

پ) فرض کنید داده زیر نیز به مجموعه داده‌های قبلی اضافه شود:

$x_4 = (2, 2), y_4 = +1$

همانند حالت‌های قبل بردارهای پشتیبان را مشخص کرده و مرز تصمیم‌گیری را بیابید. آیا مرز تصمیم‌گیری برای این حالت متفاوت از حالت (ب) خواهد بود؟ وزن بردارهای پشتیبان در این حالت نسبت به حالت قبلی کوچک‌ترند یا بزرگ‌تر؟

ت) فرض کنید داده زیر نیز به مجموعه داده‌های قبلی افزوده شود:

$x_5 = (-3,-3), y_5 = -1$

کدام یک از کرنل‌های زیر قادر به جداسازی این داده‌ها هستند؟

- کرنل خطی: $K(u, v) = u \cdot v$
- کرنل چند جمله‌ای با درجه بزرگتر مساوی ۲: $K(u, v) = (1 + u \cdot v)^n$
- کرنل گاوسی: $K(u, v) = \exp(-||u-v||/2\sigma^2)$

ث) فرض کنید دو داده جدید زیر به داده‌های قبلی افزوده شوند:

$x_6 = (+1, +1), y_6 = +1$

$x_7 = (-4,-4), y_7 = -1$

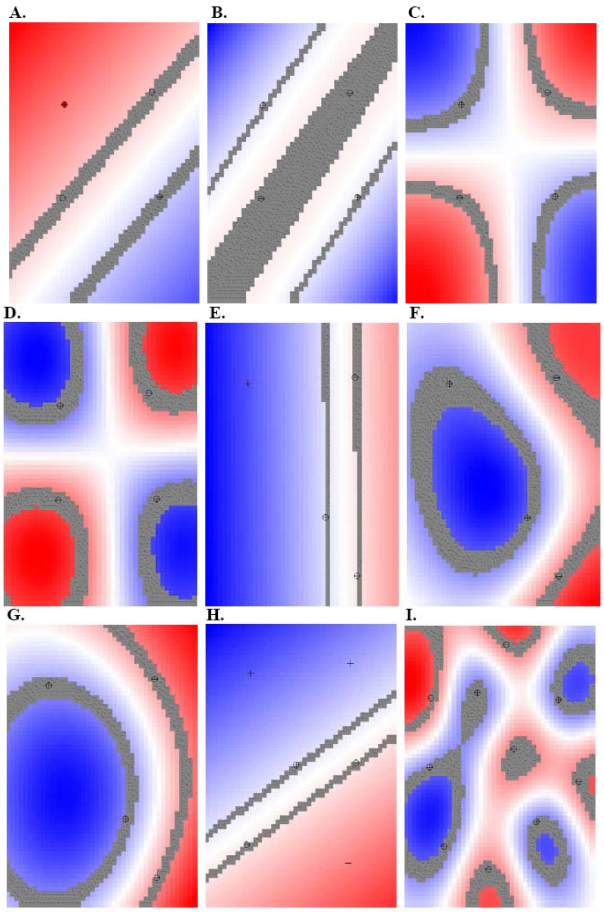
کرنلی به فرم زیر را برای این داده‌ها در نظر بگیرید:

$$K(u,v)=2||u||\,||v||$$

بردارهای پشتیبان را بیابید و کلاسه‌بند مربوطه را به فرم زیر بیابید:

$$h(x)=\begin{cases}+1 & if\ -2x_1^2+....x_1x_2+....x_2^2+....x_1+....x_2+....\geq 0\\ -1 & otherwise\end{cases}$$

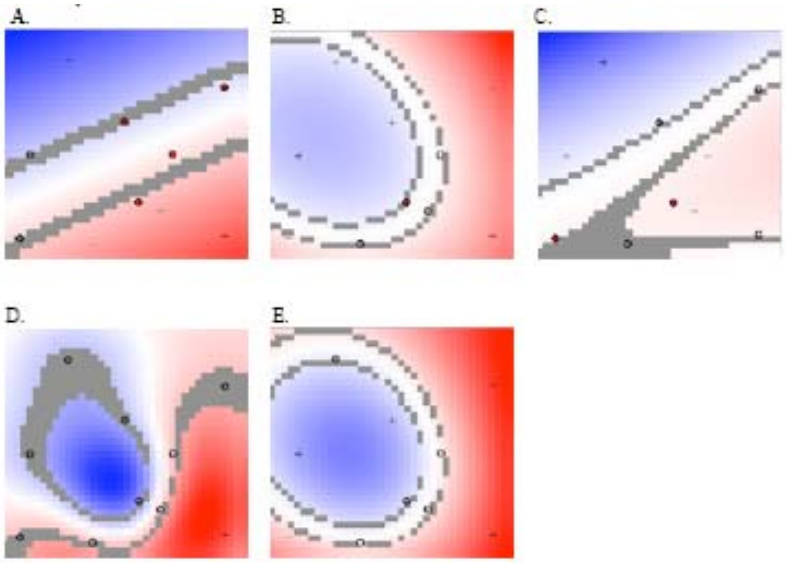
سوال ۲) برای یک مجموعه داده مشخص، ۹ دیاگرام حاصل از SVM با کرنل‌های مختلف شده در جدول زیر را در نظر بگیرید. در جدول داده شده، مشخص کنید که هر دیاگرام می‌تواند توسط کدام یک از کرنل‌ها ایجاد شود؟



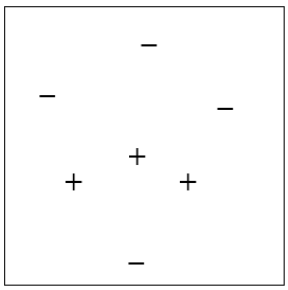
$\kappa(\upsilon_1,\upsilon_2)=(\upsilon_1.\upsilon_2)$		$\kappa(\upsilon_1,\upsilon_2)=\exp(-\frac{-\\|\upsilon_1-\upsilon_2\\|^2}{0.5})$	
$\kappa(\upsilon_1,\upsilon_2)=(1+\upsilon_1.\upsilon_2)^2$		$\kappa(\upsilon_1,\upsilon_2)=\exp(-\frac{-\\|\upsilon_1-\upsilon_2\\|^2}{0.22})$	
$\kappa(\upsilon_1,\upsilon_2)=\exp(-\frac{-\\|\upsilon_1-\upsilon_2\\|^2}{0.08})$			

سوال ۳) برای یک مجموعه داده مشخص پنج دیاگرام حاصل از SVM با کرنل‌های مختلف زیر را در نظر بگیرید. در جدول زیر مشخص کنید که هر دیاگرام می‌تواند توسط کدام یک از کرنل‌های زیر ایجاد شود؟

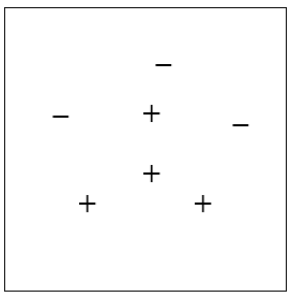
RBF, σ=0.08		Linear	
RBF, σ=0.5		Second Order Polynomial	
RBF, σ=2			



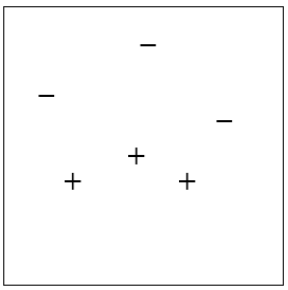
سوال ٤) دیاگرام‌های زیر مجموعه داده‌های مجزایی را نمایش می‌دهند. فرض کنید کرنل‌های $(v_1.v_2)$، $(1+v_1.v_2)^2$، RBF با $\sigma = 2$ و $\sigma = 0.1$ RBF را داشته باشیم. برای هر دیاگرام، ساده‌ترین کرنلی که تمامی داده‌ها را بدرستی کلاسه‌بندی می‌کند، تعداد بردارهای پشتیبان مرز تصمیم‌گیری و حاشیه‌های امنیت مربوطه را مشخص نمایید.



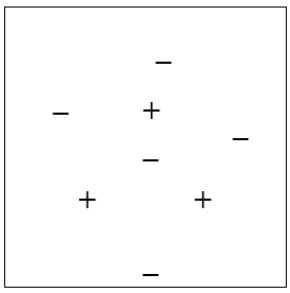
Kernel: ..... #SVs: .....



Kernel: ..... #SVs: .....

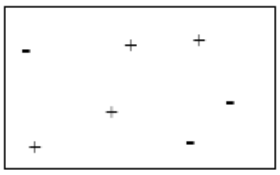


Kernel: ..... #SVs: .....

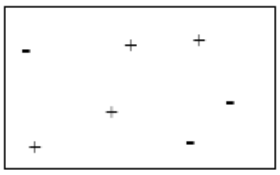


Kernel: ..... #SVs: .....

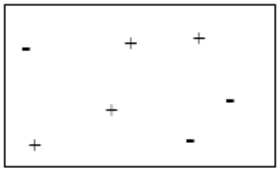
سوال ۵) دیاگرام‌های زیر مجموعه داده مشخصی را نمایش می‌دهند. برای هر دیاگرام با توجه به الگوریتم داده شـده عملیـات افراز فضا را انجام دهید:



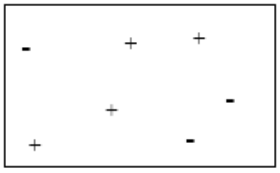
Neural Network with 1 hidden layer of 2 units



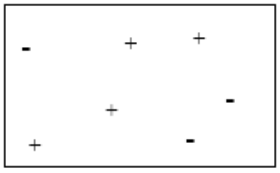
Neural Network with unlimited hidden layers/units



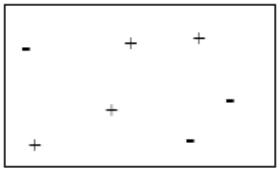
1-Nearest Neighbor



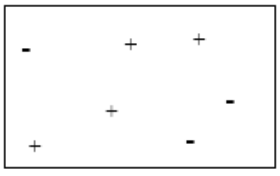
K-means (k=2)



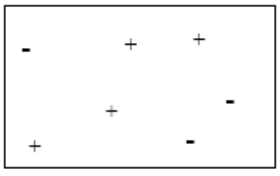
GMM with 2 Gaussian mixtures Models



GMM with 3 Gaussian Mixture Models



SVM with Linear kernel



SVM with polynomial kernel

سوال ٦) دو مدل مخفی مارکوف (HMM) دو حالته زیر را در نظر بگیرید، که هر دو حالت دارای دو خروجی ممکن A یـا B هستند.

- مدل اول:
  - احتمال‌های انتقال: $a_{11}=0.7, a_{12}=0.3, a_{21}=0.0, a_{22}=1.0$ (a_{ij} احتمال رفتن از حالت i به j)
  - احتمال‌های خروجی: $b_1(A)=0.8, b_1(B)=0.2, b_2(A)=0.4, b_2(B)=0.6$

- احتمال‌های اولیه: $\pi_1 = 0.5, \pi_2 = 0.5$

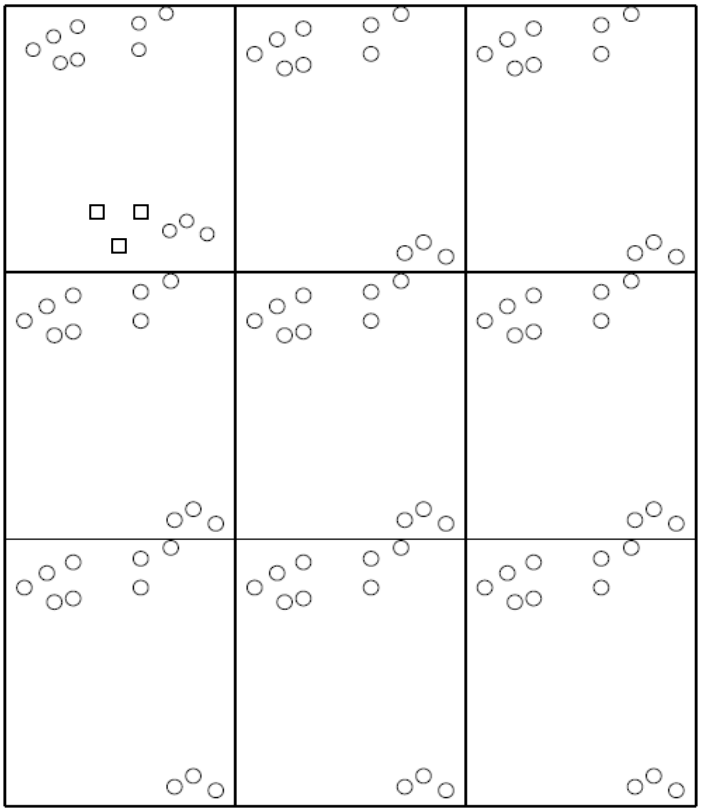
- مدل دوم:
  - احتمال‌های انتقال: $a_{11} = 0.6, a_{12} = 0.4, a_{21} = 0.0, a_{22} = 1.0$
  - احتمال‌های خروجی: $b_1(A) = 0.9, b_1(B) = 0.1, b_2(A) = 0.3, b_2(B) = 0.7$
  - احتمال‌های اولیه: $\pi_1 = 0.4, \pi_2 = 0.6$

الف) دیاگرام‌های حالت دو مدل را رسم کنید.

ب) کدام مدل با احتمال بالاتری دنباله {A, B, B} را تولید می‌کند؟

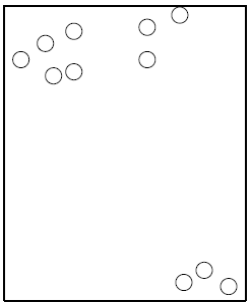
پ) برای دنباله فوق مسیر Viterbi را در هر دو مدل پیدا کنید؟ آیا نتیجه بدست آمده با نتیجه قسمت (ب) تطابق دارد؟

سوال ۷) الف) الگوریتم K-means را بصورت دستی بر روی داده‌های زیر اجرا نمائید. دایره‌ها نشان دهنده‌ی داده‌های موجود بوده و مربع‌ها، مراکز خوشه‌ها را نشان می‌دهند. مراکز ابتدایی خوشه‌ها در ابتدا بصورت زیر داده شده‌اند. مرزهایی را که توسط این مراکز پدید می‌آیند را نشان داهید (فضا به چه نحوی افراز می‌شود؟). بر روی تصاویر داده شده، مراحل الگوریتم را تا حالت همگرایی، مرحله به مرحله، نشان دهید.



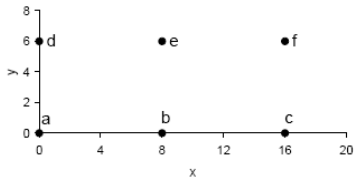
ب) فرض کنید ما توسعه‌ای از الگوریتم K-means را داریم که به جای استفاده از فاصله‌ی اقلیدسی از فاصله‌ی ماهالانوبیس استفاده می‌کند و خوشه‌های تولیدی آن به جای دایره، دارای شکل گاوسی می‌باشند (در هر مرحله از الگوریتم، پس از نسبت دادن هر کدام از داده‌ها به یکی از مراکز و تشکیل دادن خوشه‌های موقتی، از الگوریتم ML برای تخمین پارامترهای گاوسی جدید هر خوشه استفاده نموده و گاوسی‌های هر خوشه را بروز می‌کنیم). اگر این الگوریتم از همان مراکز خوشه‌ی ابتدایی بخش قبلی استفاده کند، در نهایت حدس می زنید گاوسی‌های نهایی چگونه باشند؟ بیضی‌های معرف هر مدل گاوسی را در شکل

روبرو نمایش دهید (از محاسبات دقیق خودداری نموده و شکل‌های تقریبی را حدس بزنید – کوواریانس‌های ابتدایی را که باید بصورت تصادفی در نظر گرفته شوند، شما برای راحتی کار، مطابق نیاز خود در نظر بگیرید).



پ) آیا نتایج خوشه‌بندی مراحل (الف) و (ب) فوق یکسان هستند؟ چرا؟

سوال ۸) فرض کنید مجموعه داده‌ی ۶ عضوی زیر را داشته باشیم:



S={a=(0,0), b=(8,0), c=(16,0), d=(0,6), e=(8,6), f=(16,6)}

بر روی این داده‌ها الگوریتم K-means با k=3 را می‌خواهیم اعمال کنیم. معیار ارزیابی فاصله در الگوریتم را همان فاصله اقلیدسی در نظر می‌گیریم. قبل از مطرح نمودن سوال ابتدا به دو تعریف زیر توجه نمائید:

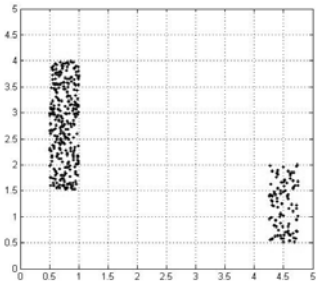
- K-Starting Configuration (K-SC): یک زیر مجموعه‌ی kتایی از S، که مراکز اولیه ما را نشان می‌دهند. مثلا {a,b,c}.
- K-partition (K-P): یک افراز از S به k زیر مجموعه‌ی غیر تهی را گوئیم. مثلا {{a,b,e}, {c,d}, {f}}.
- یک K-P پایدار خوانده می‌شود، اگر تکرار اجرای مراحل الگوریتم K-means بر روی آن باعث ایجاد تغییرات در خوشه‌های نتیجه شده نشود.

الف) تعداد 3-SCهای موجود بر روی S داده شده چند تاست؟

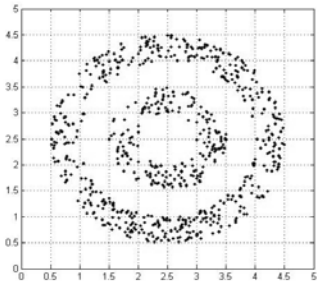
ب) جدول زیر را پر کنید:

3-P	پایدار است؟	یک 3-SC اولیه که با اجرای K-means بتواند به این 3-P منجر شود.	تعداد 3-SCهای منجر شونده به این 3-P؟
{a,b,e}, {c,d}, {f}			
{a,b}, {d,e}, {c,f}			
{a,d}, {b,e}, {c,f}			
{a}, {d}, {b,c,e,f}			
{a,b}, {d}, {c,e,f}			
{a,b,d}, {c}, {e,f}			

سوال ۹) مجموعه داده‌های داده شده در سایت درس (دو مجموعه داده A با ابعاد a1 و a2 و B با ابعـاد b1 و b2) را در نظـر بگیرید (نمای این داده‌ها را در شکل زیر مشاهده می‌کنید). برای کلاستر کردن این دو مجموعه داده کد مطلب مناسبی را ارائـه دهید و نتایج بدست آمده را گزارش دهید.



مجموعه داده A



مجموعه داده B

سوال ۱۰) فرض کنید x دارای توزیع Hat با میانگین μ باشد، یعنی:

$$p(x \mid \mu) = \begin{cases} 0 & x \le \mu - 1 \\ 1-(\mu - x) & \mu - 1 \le x \le \mu \\ 1-(x-\mu) & \mu \le x \le \mu + 1 \\ 0 & \mu + 1 \le x \end{cases}$$

فرض کنید داده‌های {1, 3, 6, 7} بوسیله ترکیبی از سه توزیع Hat تولید شده باشند و همچنین داشته باشیم:

$$P(w1)=1/2,\ P(w2)=1/4,\ P(w3)=1/4$$

سه توزیع فوق را طوری بیابید که بیشترین Likelihood را داشته باشیم.

سوال ۱۱) فرض کنید که متغیر تصادفی X دارای توزیع ترکیبی به صورت زیر باشد:

$$p_\alpha(x) = \alpha * p_1(x) + (1-\alpha) * p_2(x)$$

فرض کنید دو توزیع p_1 و p_2 برای ما شناخته شده باشند و فقط متغیر α برای ما ناشناخته باشد. فرض کنید n نمونه داده i.i.d $\{x_1, x_2, ..., x_n\}$ از توزیع X را داشته باشیم. یک الگوریتم EM برای تخمین α ارائه دهید (قـدم‌هـای E و M را در الگـوریتم خود بصورت دقیق مشخص نمائید).

سوال ۱۲) فرض کنید که $Y_1 \sim exp(1/\theta_1)$ و $Y_2 \sim exp(1/\theta_2)$ مستقل از هم باشند و $\theta_1 <> \theta_2$. فرض کنیـد n نمونـه داده i.i.d $\{x_1, x_2, ..., x_n\}$ از توزیع $X=Y_1+Y_2$ را داشته باشیم.

الف) توزیع X را بیابید.

نکته ۱: چگالی Y_1 برابر است با $f_{\theta_1}(y) = \theta_1 exp(-\theta_1 y)$. برای Y_2 هم چگالی به همین صورت می‌باشد.

نکته ۲: ابتدا CDF را برای X بدست آورید، یعنی $F(x) = P(Y_1 + Y_2 < x) = \int_0^x \int_0^{x-y_1} f_{\theta_1}(y_1) f_{\theta_2}(y_2) dy_2 dy_1$

ب) برای تخمین MLE پارامترهای θ_1 و θ_2 الگوریتم EM مناسب ارائه کنید. قدم‌های E و M و عبارات بروزرسانی مربوطـه را بصورت دقیق مشخص نمائید.